MVA2002 IAPR Workshop on Machine Vision Applications, Dec. 11 - 13, 2002, Nara - ken New Public Hall, Nara, Japan

13—30

A Model for Calculating Saliency from Both Input Images and Memory

Toshio Endoh[1] Makoto Goto[2]
IT Media Laboratories
Fujitsu Laboratories Ltd.

Takashi Toriu[3]
Graduate School of Engineering
Osaka City University

Abstract

As a first step to implement human functions related to visual attention in computer vision, we developed a computational model for calculating the saliency map of an input image. This model is based on the features obtained through K-L transformation of many images, and it explains the memory-related asymmetrical effect in visual search: the search time for a familiar object among unfamiliar ones is longer than the search time for an unfamiliar object among familiar ones. We conducted a psychophysical experiment in which two subjects searched for "rotated 2"s and "rotated 5"s for 10 minutes a day throughout a six-month period and found that visual perceptual learning yielded an asymmetrical effect in visual search.

1 Introduction

Human vision does not process input information uniformly. Visual attention focuses on a limited area (attended area) in the field of view and then shifts from one area to another, depending on the situation and task. Allocating computational resources intensively to the attended area enables rapid reaction. The aim of this paper is to suggest a model that can quantitatively calculate the degree to which attention is directed to a certain area.

Visual attention has been studied through psychophysical experiments of visual search phenomena for years. In those experiments, subjects had to detect a target among several distractors . Treisman and Gelade [1] reported the results of those experiments. The reaction time was short and remained constant independent of the number of distractors (pop out) in tasks where the target was distinguishable with respect to one feature, such as the task of searching for a green "T" among brown "T"s, or the task of searching for a green "T" among green "X"s. In contrast, the reaction time increased in proportion to the number of distractors in tasks where the target was defined with respect to a combination of two features, such as the task of searching for a green "T" among brown "T"s and green "X"s.

Treisman and Gelade [1] developed the Feature Integration Theory (FIT) through those experiments. According to FIT, the process of early vision is subdivided into two processes: the pre-attentive process in which primitive features of a visual stimulus are processed in parallel, and the attentive process in which attention is focused on a certain area where several primitives, such as color and orientation, are integrated.

Koch and Ullman [2] suggested a neural network model to describe visual selective attention. In their model, a so-called Winner-Takes-All network calculates the location into which attention should be shifted. Later, Itti, and Koch [3] further developed this model and suggested a revised model that calculates a "saliency map", which topographically codes for local conspicuity over the entire complex scene. They tested the validity of this model by applying it to real images.

Wang, Cavanagh, and Green [4] found a psychological phenomenon that suggests parallel processing in the search for an unfamiliar target among familiar distractors, and serial processing in the search for a familiar target among unfamiliar distractors. Although this asymmetrical effect in visual search implies that the degree to which attention can be easily directed to a certain area of the image is influenced by visual experience, the model in the paper [3] does not cover this effect.

In this paper, we suggest a computational model that calculates the saliency map of an input image based not only on the input image information but also on the memory acquired through learning. This model determines primitives by using K-L transformation for many learning patterns, and it can simulate the asymmetry of visual search affected by memory.

In the next chapter, we describe the ideas behind our model and an algorithm used in the model. In Chapter 3, we show the results of an experiment using real images in which we found that an asymmetrical visual search can be simulated. In chapter 4, we show the results of an experiment in which we found that a six-month-long visual experience yielded an asymmetrical effect in visual search.

2 The model for calculating the saliency

We hypothesized that attention can be easily attracted to areas having features that are different from those in the surrounding areas because the part of an image that has features that are different from those in other parts pops out, and that attention can be attracted to areas that have unfamiliar features because the search time for an unfamiliar target among familiar distractors is critically short.

[1] Address: 4-1-1 Kamikodanaka, Nakahara-ku, Kawasaki 211-8588 Japan, E-mail: endow@flab.fujitsu.co.jp

[2] Address: 4-1-1 Kamikodanaka, Nakahara-ku, Kawasaki 211-8588 Japan, E-mail: mgoto@labs.fujitsu.com

[3] Address: 3-3-138 Sugimoto, Sumiyoshi-ku, Osaka 558-8588, Japan. E-mail: toriu@info.eng.osaka-cu.ac.jp

Based on these facts, we devised a model for calculating the saliency map of an input image. The model consists of a process of learning and a process of searching for unknown images. The configuration of our visual search model is shown in Figure 1.

In the learning process, for each pixel of a learning image S(x,y), autocorrelation

$$C(a,b;x,y)=\sum_{x'=x-M}^{x+M}\sum_{y'=y-M}^{y+M}S(x',y')S(x'+a,y'+b) \quad (1)$$

is first calculated as a shift invariant feature, where a and b are integers ranging from –N to N. This process extracts a (2N + 1)(2N + 1)-dimensional feature vector for each pixel in the learning image. We calculate the eigenvectors of the covariance matrix made from many feature vectors obtained in the way described above, and extract the basis of principal components from the largest eigenvalue up to a certain (e.g. 98%) cumulative contribution (K-L transform). The basis is saved as information that defines the feature that should be extracted in the search process. Thus, features extracted in a certain period of search time vary depending on the visual experience.

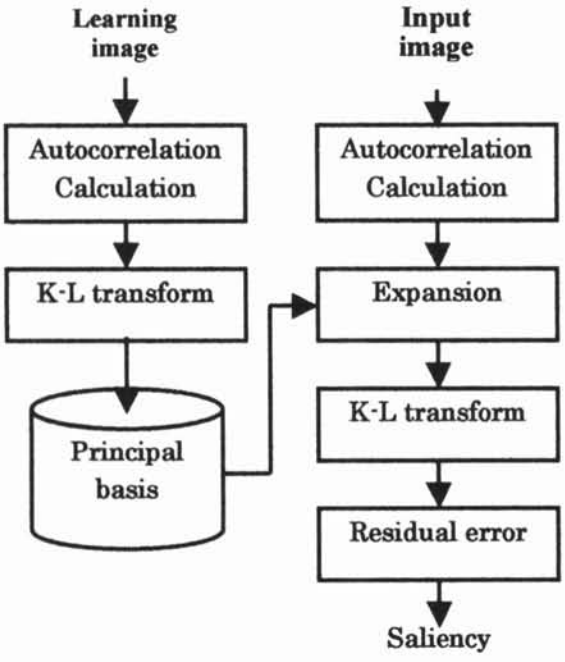

Figure 1. Model configuration

In the search process, a small window is scanned in the input image, and autocorrelation is calculated according to Equation 1. A feature vector is then extracted for each pixel, just as in the learning process. Then the vectors are expanded on the "memorized" basis, and their coefficients are obtained. These coefficients represent the local features for each part of the image. The process is shown in Figure 2. In the feature vector space, a subspace is defined by the principal basis that was extracted in the learning process. The feature vectors extracted for each pixel are projected onto this subspace. The projections are local features.

Then we calculate the eigenvectors of the covariance matrix made from all of the projected feature vectors, and extract the basis of principal components from the largest eigenvalue up to a certain (e.g. 98%) cumulative contribution (K-L transform). Finally, the feature vector extracted at each point is projected onto the subspace spanned by the principal basis, and the distance between the feature vectors and the projected vectors (i.e., residual error) is output as the saliency at the point, as shown in Figure 3.

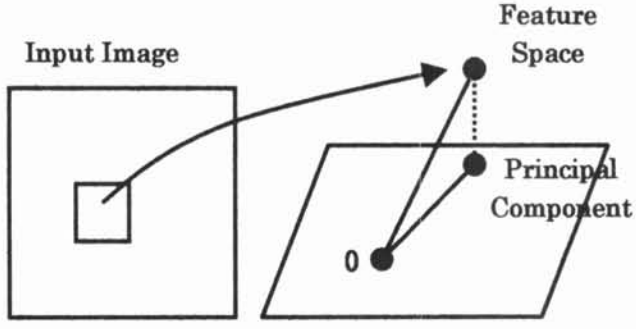

Figure 2 Feature projection

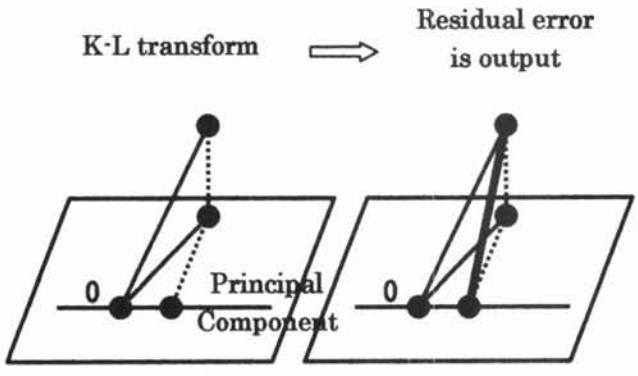

Figure 3 Calculating residual error

3 Experimental results

We conducted an experiment to evaluate the validity of our model, using images in Figure 4 (a)–(e) (left). The images on the right show the calculated saliency. The intensity of the images represents the saliency. The pixel values were normalized so as the maximum value to be constant.

In experiment (a), a thousand scenery images were used as learning images, and in experiments (b) through (e), images in which 26 capital letters in the English alphabet and 10 Arabic numerals were arranged randomly were used as learning images. The size of the letters and numerals in the learning images was the same as the size of those in the images used in the search process.

In (a), the calculated saliency was large in the area

where the orientation of one line segment was different from that of the other line segments, which confirmed the psychological phenomenon of "popping out". This result can be explained as follows: line segments oriented 135 degrees yield a small residual error in expanding the feature by the basis in search time, because there are so many 135-deg segments in the search period that the basis has enough information to represent 135-deg segments. In contrast, line segments oriented 45 degrees yield a larger residual error.

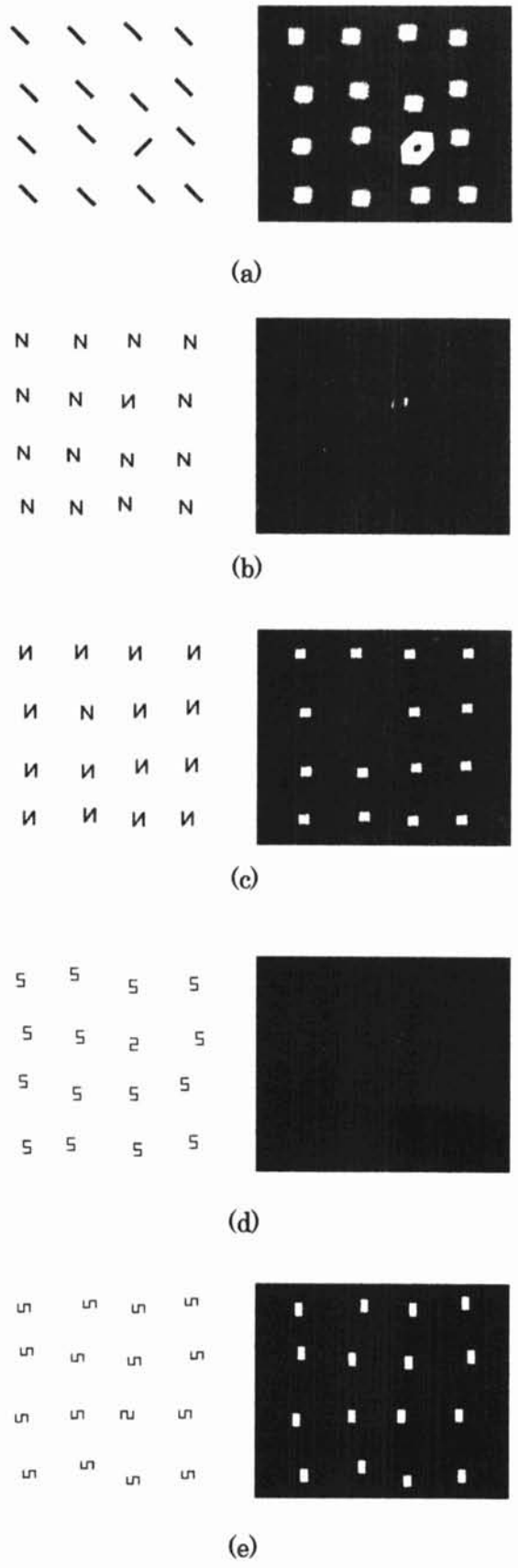

Figure 4. Calculating saliency

In (b), the saliency was large in the area with a "reversed N", which confirms the psychological results showing that the search time required to find an unfamiliar target among familiar distractors is short. This is because "reversed N"s appear only in the search period and do not appear in the learning period. In contrast, in (c), which shows the reverse pattern of (b), the saliency was not large in the area where the target was, which confirms the psychological results showing that the search time required to find a familiar target among unfamiliar distractors is long. The reason why the area with the target did not have large saliency in spite of the fact that the target differed from the other stimuli is that since there were no "reversed N"s in the learning patterns, the memorized basis did not have enough information to represent a "reversed N".

In (d), the saliency was large in the area around "upright 2", the target that differed from the other stimuli, while in (e), which was a rotated pattern of (d), the saliency in the area around "rotated 2" was not large. This is because the memorized basis did not have enough information to represent "rotated 2"s and "rotated 5"s since they did not appear in the learning images.

4 Experiment yielding asymmetry

The search time for the unfamiliar target "rotated 2" among unfamiliar "rotated 5"s was quite long, as shown in (e) in the last chapter. Similarly, the search time for a "rotated 5" among "rotated 2"s was long. But what will the result be like, if, for example, "rotated 2"s are presented to the subjects in the training? To investigate the problem, we conducted the following psychological experiment, using two subjects (TT, MG).

The tasks for the subjects included a learning task, in which either "rotated 2"s or "rotated 5"s were presented to the subjects, and a test task, in which the subjects searched for a "rotated 2" or a "rotated 5" among "rotated 5"s and "rotated 2"s, respectively. Three sessions for the test task were conducted: one session before the six-month training period, one in the middle of the training period, and one after the training period. In the sessions, the subjects had to

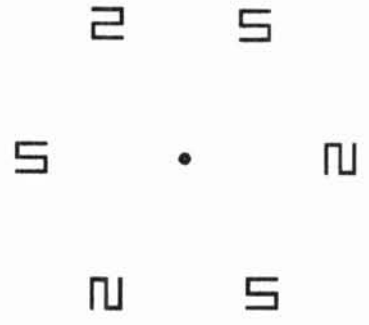

Figure 5. An arrangement of stimuli

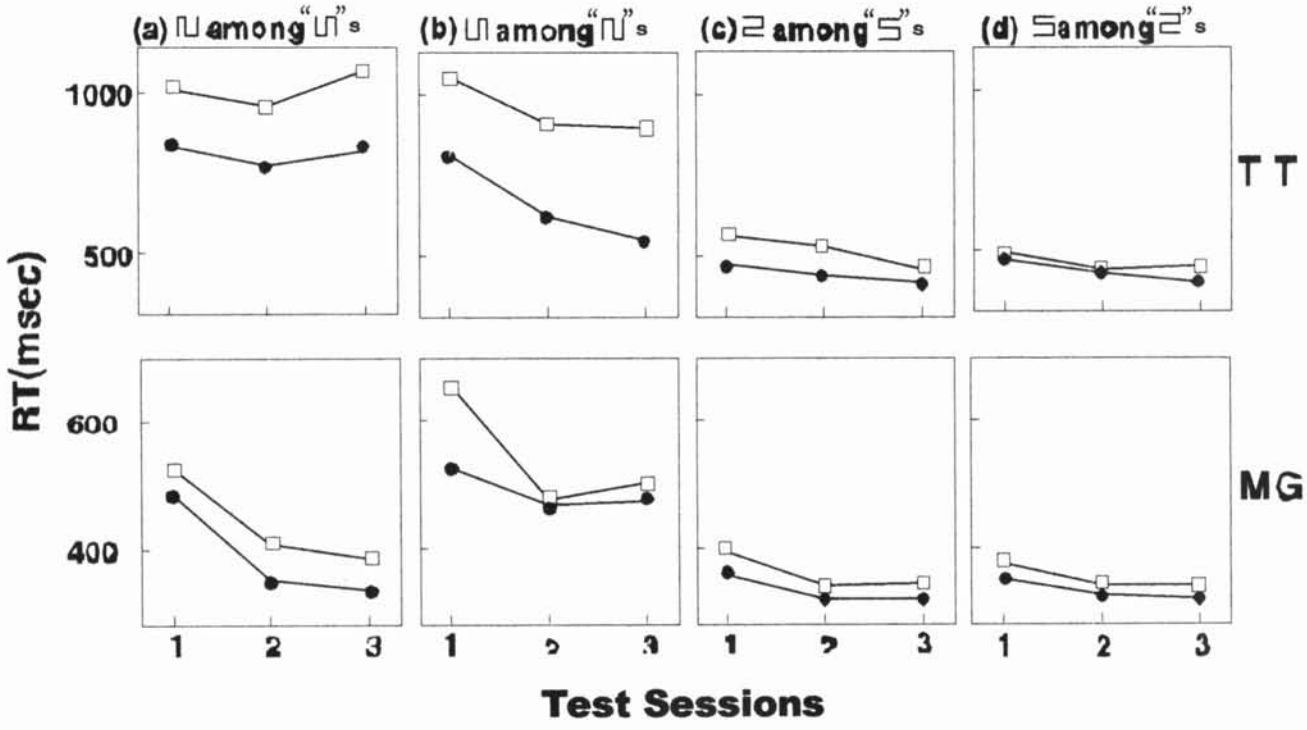

Figure 6. Changes in the mean reaction time

perform another task for comparison, which was to search for an "upright 2" or an "upright 5" among "upright 5"s and "upright 2"s, respectively.

In the learning period, subject TT searched for a target among three kinds of stimuli, namely, "rotated 2"s, "upright 2"s, and "upright 5"s, on a background of stimuli of the other two kinds. An example of this arrangement is shown in Figure 5. Subject MG searched for a target among three kinds of stimuli, namely "rotated 5"s, "upright 2"s, and "upright 5"s, on a background of stimuli of the other two kinds. The training sessions were conducted for about 10 minutes a day, during approximately six months.

The changes in the mean reaction time in the three test sessions are shown in Figure 6. The filled circles represent the time it took the subjects to find the target in target-present trials, and the open squares represent the time it took the subjects to push a button to indicate that the target was absent in target-absent trials.

When both the target and the distractors were familiar, the reaction time was short. In contrast, when the target was familiar but the distractors were not, the reaction time decreased in proportion to the degree to which the subjects were familiar with the stimuli.

5 Conclusions

We described a model for calculating the saliency map of an input image, which can explain the asymmetrical effect in visual search affected by memory. The model embodies the memory effect by using a principal basis obtained from K-L expansion of learning images as features during the search. We found that the asymmetry in visual search appeared again after six months of perceptual learning. Modeling the contextual effect, such as that of the situation in which the search process takes place or a task is performed, is the topic of our future work.

References

[1] A. M. Treisman and G. Gelade, "A Feature-Integration Theory of Attention," Cognitive Psychology Vol. 12, No. 1, pp. 97-136, 1980.

[2] C. Koch, and S. Ullman, "Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry," Human Neurobiology, Vol. 4, pp. 219-227, 1985.

[3] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," IEEE Trans. PAMI, Vol. 20, No. 11, pp. 1254-1259, 1998.

[4] Q. Wang, and P. Cavanagh and M. Green, "Familiarity and pop-out in visual search.," Perception & Psychophysics, Vol. 56, p.495-500, 1994.